

A two-queue Markovian polling system with two priority levels in the first queue

Chu Yuqing*, Liu Zaiming, Wu Jinbiao

*Department of Mathematics and Statistics, Central South University, Changsha,
Hunan 410083, PR China*

Abstract

In this paper, we deal with a two-queue polling system attended by a single server. The server visits the queues according to a Markovian routing mechanism. There are two-class customers in the first queue. Customers of each queue are served in the exhaustive discipline. For this model, we study the joint number of customers, the cycle time and the waiting times. We are also devoted to deriving the exact asymptotics for the scaled delay in the heavy-traffic scenario. In addition, the scaled delays with large switch-over times are discussed for the first time in the Markovian polling systems. Finally, we provide some simulations to support the asymptotic results.

Keywords:

Polling System, Markovian Routing, Priority, Heavy-traffic, Large Switch-over Time, Simulation

Mathematics Subject Classification (2010): 60K25

1. Introduction

In this paper, we analyze a two-queue Markovian polling system. The random routing discipline is widely used in computer-communication systems with random access schemes and production systems with multi-type products, such as ALOHA and CSMA-CA (Carrier-Sense Multiple-Access Collision-Avoidance) algorithms. In the modelling of cellular data services, access is randomly assigned to the multitude of users in a cell. More specifically, a time-slot (representing the right for transmission) is assigned to the user with the highest signal-to-noise

*Corresponding author

Email address: chuyuqing@csu.edu.cn (Chu Yuqing), math_lzm@csu.edu.cn (Liu Zaiming), wujinbiao@csu.edu.cn (Wu Jinbiao)

ratio among all users in a cell, which is so-called opportunistic scheduling. The opportunistic scheduling aims to improve the bandwidth efficiency by utilizing fading and shadowing of cellular users within a single cell.

Although there is quite an extensive amount of literature on polling systems, only very few papers concern with Markovian polling systems. The few papers that can be found include [6], where a pseudo-conservation law for mean waiting times was derived. Later, a more general Markovian polling system with a dependent routing probabilities was studied in [14]. More recently, Dorsman considered the applications to wireless random-access networks and analyzed some optimizing problems in [7]. He also considered the case of two queues specially in [8], where the heavy-traffic behavior was investigated with the descendent set approach (DSA) for the first time.

Here we introduce priority policy to differentiate between high priority traffic, like streaming multimedia, and low priority traffic, like web browsing, to improve the QoS (Quality of Service) standard, just like the application of the threshold priority policy in [13, 12]. Only very few papers treat priorities in polling systems. Boon first proved the queue lengths at polling epochs do not depend on the service order and hence gave their Probability Generating Functions (PGFs) in priority polling systems in [3]. Besides, he applied the notion of delay-cycles, introduced in [11], and presented the LSTs of the waiting times. The model was further extended with multiple priority levels in [2] and with mixed gated/exhaustive service discipline in [1]. Here, we apply the same methodologies to the Markovian polling systems.

Apart from the aforementioned performances, we also discuss the limiting delays with large switch-over times. The case of large deterministic switch-over times is explored by using the DSA in combination with the Strong Law of Large Numbers for Renewal Processes, while the delay with large general switch-over times is studied under heavy traffic.

The remainder paper is organized as follows. We first give a detailed description of the model in Section 2 and then present the analysis of the non-priority polling system as the preliminary work in Section 3. Section 4 is dedicated to the derivation of the cycle time, the joint queue lengths at polling epochs and at an arbitrary time, and the waiting times. The scaled delays in the heavy-traffic and with large switch-over times are explored in Section 5 and Section 6 respectively. In Section 7, simulations are undertaken to test the validity of limiting theorems. We finally conclude the whole procedure and propose some topics for further research in Section 8.

2. Model description

We consider a single server Markovian polling model with two queues Q_1 and Q_2 . Customers arriving at Q_2 are referred to type-2 customers, while Q_1 contains two types of customers: type- H and type- L customers. Type- H customers have non-preemptive priority over type- L customers. The buffer capacity of each queue is infinite. Each type of customers are served in FCFS discipline. Type- i customers arrive independently according to a Poisson process with rate λ_i and the service times B_i are mutually independent with LST $\tilde{B}_i(s)$ and mean $\mathbb{E}B_i = \frac{1}{\mu_i}$, $i = H, L, 2$. Then the traffic load of Q_i equals $\rho_i = \lambda_i \mathbb{E}B_i$ ($i = 1, 2$) with $\lambda_1 = \lambda_H + \lambda_L$ and $B_1 = \frac{\lambda_H}{\lambda_1} B_H + \frac{\lambda_L}{\lambda_1} B_L$. Let $\rho_i = \frac{\lambda_i}{\mu_i}$ ($i = H, L$). Therefore, $\rho_1 = \rho_H + \rho_L$. It is assumed that each queue is served exhaustively. We also assume the total traffic load $\rho = \rho_1 + \rho_2 < 1$, according to the conclusion in [9], which implies the system is stable.

The Markovian routing mechanism means that the next queue polled will be determined by a discrete time Markov chain. For convenience, we denote the Markov chain as $M = \{d_n, n \geq 0\}$ with state space $I = \{1, 2\}$, where $d_n = i$ means the n th polled queue after $t = 0$ is Q_i . It is assumed that once the server completes service of Q_i , it begins a switch-over time S_{ii} for another poll at Q_i with probability $p_i < 1$ and begins a switch-over time S_{ij} for a poll at Q_j with probability $1 - p_i$ ($i = 1, 2$). Define

$$\begin{aligned}\pi_i &= \lim_{n \rightarrow \infty} Pr\{d_n = i\}, \quad i \in I, \\ r_{ij} &= Pr\{d_n = i | d_{n+1} = j\}, \quad i, j \in I, \quad n = 0, 1, \dots\end{aligned}$$

With the notations, the average duration of an arbitrary switch-over time is given by $\sigma = \sum_{i=1}^2 \sum_{j=1}^2 r_{ij} \pi_j E[S_{ij}]$. And it is easy to compute

$$\begin{aligned}\pi_1 &= \frac{1 - p_2}{2 - p_1 - p_2}, & \pi_2 &= \frac{1 - p_1}{2 - p_1 - p_2}, \\ r_{11} &= p_1, & r_{12} &= 1 - p_2, \\ r_{21} &= 1 - p_1, & r_{22} &= p_2.\end{aligned}$$

3. Preliminary

To avoid the tedious calculations, we follow the same idea in [3] to give the joint number of customers at polling epochs. This section contains some preliminaries concerning the analysis of non-priority polling system. “Non-priority” means that we combine type- H and type- L customers into general type-1 customers with arriving rate λ_1 and service times B_1 . For compactness, we omit the derivation process, which can be referred to [8]. It is noted that throughout the paper, for the random variable X , $\tilde{X}(\cdot)$ denotes its PGF or LST.

Let K_{ij} denote the number of type- j customers arriving at Q_i during a busy period. Then its PGF is given by

$$\tilde{K}_{ij}(z) = \tilde{\theta}_i(\lambda_j(1-z)), \quad i, j = 1, 2,$$

where θ_i denotes a busy period in an $M/G/1$ queue with arrival rate λ_i and service rate μ_i .

Let $M_{ij}^{(k)}$ be the number of arriving type- k customers during a switch-over time S_{ij} . Then the PGF of the joint distribution of $M_{ij}^{(k)}$ is given by

$$\widetilde{M}_{ij}(z_1, z_2) = \mathbb{E}[z_1^{M_{ij}^{(1)}} z_2^{M_{ij}^{(2)}}] = \tilde{S}_{ij}(\lambda_1(1-z_1) + \lambda_2(1-z_2)), \quad i, j = 1, 2.$$

Along the arguments in [8], the PGFs of the joint queue lengths at the polling epochs at Q_i , denoted by $\tilde{V}_{bi}(z_1, z_2)$, are given by

$$\tilde{V}_{b1}(z_1, z_2) = r_{11}\widetilde{M}_{11}(z_1, z_2) \prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2)) + r_{21}\widetilde{M}_{21}(z_1, z_2) \prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)), \quad (1)$$

$$\tilde{V}_{b2}(z_1, z_2) = r_{12}\widetilde{M}_{12}(z_1, z_2) \prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2)) + r_{22}\widetilde{M}_{22}(z_1, z_2) \prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)), \quad (2)$$

where

$$\begin{aligned} f_1(z_2) &= \tilde{K}_{21}(\tilde{K}_{12}(z_2)), & f_2(z_1) &= \tilde{K}_{12}(\tilde{K}_{21}(z_1)), \\ f_1^{(0)}(z_2) &= z_2, & f_1^{(j)}(z_2) &= f_1(f_1^{(j-1)}(z_2)), \quad j = 1, 2, \dots, \\ f_2^{(0)}(z_1) &= z_1, & f_2^{(j)}(z_1) &= f_2(f_2^{(j-1)}(z_1)), \quad j = 1, 2, \dots, \\ a_1(z_2) &= \frac{r_{21}\widetilde{M}_{21}(\tilde{K}_{12}(z_2), z_2)}{1 - r_{11}\widetilde{M}_{11}(\tilde{K}_{12}(z_2), z_2)} \frac{r_{12}\widetilde{M}_{12}(\tilde{K}_{12}(z_2), f_1(z_2))}{1 - r_{22}\widetilde{M}_{22}(\tilde{K}_{12}(z_2), f_1(z_2))}, \\ a_2(z_1) &= \frac{r_{12}\widetilde{M}_{12}(z_1, \tilde{K}_{21}(z_1))}{1 - r_{22}\widetilde{M}_{22}(z_1, \tilde{K}_{21}(z_1))} \frac{r_{21}\widetilde{M}_{21}(f_2(z_1), \tilde{K}_{21}(z_1))}{1 - r_{11}\widetilde{M}_{11}(f_2(z_1), \tilde{K}_{21}(z_1))}. \end{aligned}$$

Substituting $z_2 = 1$ in (1), we get

$$\tilde{V}_{b1}(z_1, 1) = r_{11}\widetilde{M}_{11}(z_1, 1) + r_{21}\tilde{G}_1(z_1), \quad (3)$$

where

$$\tilde{G}_1(z) = \widetilde{M}_{21}(z, 1) \prod_{j=0}^{\infty} a_2(f_2^{(j)}(z)), \quad (4)$$

Here $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z))$ can be interpreted as the total number of currently present type-1 customers whose descendants arrive during all the *inter visit-end periods* of Q_2 (see [8]) in the past. Besides, $\tilde{G}_1(z)$ can be computed by the DSA as follows.

$$\tilde{G}_1(z) = \tilde{H}_1(z) \frac{1 - r_{11}\widetilde{M}_{11}(z, 1)}{r_{21}}, \quad (5)$$

where H_1 satisfies

$$\begin{aligned} \tilde{H}_1(z) = & \prod_{c=0}^{\infty} \left[\tilde{R}_1 \left(\lambda_1(1 - \tilde{A}_{1,c-1}(z)) + \lambda_2(1 - \tilde{A}_{2,c}(z)) \right) \right. \\ & \left. \times \tilde{R}_2 \left(\lambda_1(1 - \tilde{A}_{1,c-1}(z)) + \lambda_2(1 - \tilde{A}_{2,c-1}(z)) \right) \right]. \end{aligned} \quad (6)$$

with

$$\begin{aligned} \tilde{R}_1(s) &= \tilde{S}_{12}(s) \frac{r_{12}}{1 - r_{22}\tilde{S}_{22}(s)}, \\ \tilde{R}_2(s) &= \tilde{S}_{21}(s) \frac{r_{21}}{1 - r_{11}\tilde{S}_{11}(s)}, \\ \tilde{A}_{1,c}(z) &= \tilde{\theta}_1(\lambda_2(1 - \tilde{A}_{2,c}(z))) = \tilde{K}_{12}(\tilde{A}_{2,c}(z)), & \tilde{A}_{1,-1}(z) &= z, \\ \tilde{A}_{2,c}(z) &= \tilde{\theta}_2(\lambda_1(1 - \tilde{A}_{1,c-1}(z))) = \tilde{K}_{21}(\tilde{A}_{1,c-1}(z)), & \tilde{A}_{2,-1}(z) &= 1. \end{aligned}$$

Let V_i denote the visit time at Q_i , I_i denote the intervisit time of Q_i and C_i denote the cycle time starting with a visit beginning at Q_i ($i=1,2$). For exhaustive service discipline, it is evident that

$$\tilde{I}_1(s) = \tilde{V}_{b_1}(1 - \frac{s}{\lambda_1}, 1), \quad \tilde{C}_1(s) = \tilde{V}_{b_1}(\tilde{\theta}_1(s) - \frac{s}{\lambda_1}, 1), \quad (7)$$

$$\tilde{I}_2(s) = \tilde{V}_{b_2}(1, 1 - \frac{s}{\lambda_2}), \quad \tilde{C}_2(s) = \tilde{V}_{b_2}(1, \tilde{\theta}_2(s) - \frac{s}{\lambda_2}). \quad (8)$$

In addition, with the balance arguments, the first moment can be obtained more easily. Here we just mention the results that $\mathbb{E}C_i = \frac{\sigma}{\pi_i(1-\rho)}$ and $\mathbb{E}I_i = \frac{\sigma(1-\rho_i)}{\pi_i(1-\rho)}$ ($i = 1, 2$). See [7] for more details.

4. Analysis

For a random variable X , we denote its PGF in priority polling system by $\tilde{X}(\cdot, \cdot, \cdot)$ and let $\tilde{X}(\cdot, \cdot)$ be the analogous quantities for non-priority system.

4.1. Joint number of customers at polling epochs and cycle time

Since the joint number of customers at polling epochs and the cycle time do not depend on the order of service, they remain the same as that in the non-priority system, which is rigorously proved in Lemma 1 in [3]. Hence, we have the following theorems.

Theorem 1. *The PGFs of the joint number of customers at each polling epochs are given by*

$$\tilde{V}_{b_i}(z_H, z_L, z_2) = \tilde{V}_{b_i}(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1}, z_2), \quad i = 1, 2.$$

In addition, (7) and (8) are still valid for the LSTs of the cycle times and intervisit times.

4.2. Waiting time

From the viewpoint of a type- H customer, the polling system is an $M/G/1$ queue with multiple vacations: the intervisit time I_1 and the service time of a type- L customer B_L . Hence, we can use the concept of delay-cycles, introduced in [11], a methodology which is devoted to deriving the waiting times in $M/G/1$ queue with priorities and vacations, to compute the waiting time of a type- H customer.

The delay-cycle for a tagged type- H customer is a cycle that starts with a certain initial delay that no type- H customer is waiting in line and terminates at the moment that no type- H customers are present in the system. For simplicity, the delay-cycle with initial delay I_1 is denoted by I_H cycle and the delay-cycle with initial delay B_L is denoted by L_H cycle. It is noted that the arrival of a tagged type- H customer must take place within some delay-cycle. The fraction of the time that the system is in an L_H cycle equals $\lambda_L \cdot \frac{\mathbb{E}B_L}{1-\rho_H} = \frac{\rho_L}{1-\rho_H}$ and the fraction in an I_H cycle equals $1 - \frac{\rho_L}{1-\rho_H} = \frac{1-\rho_L}{1-\rho_H}$. By the Fuhrmann-Cooper decomposition Theorem in [10], the LST of the waiting time of a type- H customer is expressed by

$$\begin{aligned}\widetilde{W}_H(s) &= \frac{(1-\rho_H)s}{s-\lambda_H(1-\widetilde{B}_H(s))} \left[\frac{1-\rho_L}{1-\rho_H} \frac{1-\widetilde{I}_1(s)}{s\mathbb{E}I_1} + \frac{\rho_L}{1-\rho_H} \frac{1-\widetilde{B}_L(s)}{s\mathbb{E}B_L} \right] \\ &= \frac{\pi_1(1-\rho) \left[1 - \widetilde{V}_{b_1}(1 - \frac{s}{\lambda_1}, 1) \right]}{\sigma[s - \lambda_H(1 - \widetilde{B}_H(s))]} + \frac{\rho_L \left[1 - \widetilde{B}_L(s) \right]}{[s - \lambda_H(1 - \widetilde{B}_H(s))]\mathbb{E}B_L}.\end{aligned}$$

Now we introduce a *completion time* with LST of

$$\widetilde{B}_{L'}(s) = \widetilde{B}_L(s + \lambda_H(1 - \widetilde{\theta}_H(s))), \quad (9)$$

where θ_H denotes a busy period in an $M/G/1$ queue with arrival rate λ_H and service rate μ_H , which is actually the service time of a type- L customer, plus the service time of its type- H descendants. Then for a type- L customer, the polling system is an $M/G/1$ queue with vacation $I_{1'}$ of LST:

$$\widetilde{I}_{1'}(s) = \widetilde{I}_1(s + \lambda_H(1 - \widetilde{\theta}_H(s))).$$

Customers arrive with rate λ_L and have service time $B_{L'}$. The traffic load equals $\rho_{L'} = \frac{\rho_L}{1-\rho_H}$. By the Fuhrmann-Cooper decomposition Theorem in [10], it is readily to obtain

$$\begin{aligned}\widetilde{W}_L(s) &= \frac{(1-\rho_{L'})s}{s-\lambda_L(1-\widetilde{B}_{L'}(s))} \frac{1-\widetilde{I}_{1'}(s)}{s\mathbb{E}I_{1'}} \\ &= \frac{\pi_1(1-\rho)}{\sigma[s - \lambda_L(1 - \widetilde{B}_{L'}(s))]} \left[1 - \widetilde{V}_{b_1}(\widetilde{\theta}_H(s), 1 - \frac{s}{\lambda_L}, 1) \right].\end{aligned}$$

Note that the waiting time of a type-2 customer remains the same as that in the non-priority system. Hence, applying the conclusion in [5] leads to

$$\widetilde{W}_2(s) = \frac{\pi_2(1-\rho)}{\sigma[s - \lambda_2(1 - \widetilde{B}_2(s))]} \left[1 - \widetilde{V}_{b_2}(1, 1, 1 - \frac{s}{\lambda_2}) \right].$$

Now all the derivations in this subsection can be summarized as:

Theorem 2. *The LST of the waiting times of each type customer are expressed by*

$$\widetilde{W}_H(s) = \frac{\pi_1(1-\rho) \left[1 - \widetilde{V}_{b_1}(1 - \frac{s}{\lambda_1}, 1) \right]}{\sigma[s - \lambda_H(1 - \widetilde{B}_H(s))]} + \frac{\lambda_L \left[1 - \widetilde{B}_L(s) \right]}{[s - \lambda_H(1 - \widetilde{B}_H(s))]}, \quad (10)$$

$$\widetilde{W}_L(s) = \frac{\pi_1(1-\rho)}{\sigma[s - \lambda_L(1 - \widetilde{B}_{L'}(s))]} \left[1 - \widetilde{V}_{b_1}(\widetilde{\theta}_H(s), 1 - \frac{s}{\lambda_L}, 1) \right], \quad (11)$$

$$\widetilde{W}_2(s) = \frac{\pi_2(1-\rho)}{\sigma[s - \lambda_2(1 - \widetilde{B}_2(s))]} \left[1 - \widetilde{V}_{b_2}(1, 1, 1 - \frac{s}{\lambda_2}) \right]. \quad (12)$$

4.3. Joint number of customers at an arbitrary time

Unlike the joint number of customers at polling epochs, the joint number of customers at an arbitrary time does depend on the service order. Now we turn this priority model into an equivalent non-priority model by dividing Q_1 into two queues Q_H and Q_L . To distinguish from the definitions of type- H and type- L customers, we take customers at Q_H as type- H' customers and customers at Q_L as type- L' customers. Type- L' customers arrive with intensity λ_L , while type- H' customers belong to *smart customers* described in [4], that if the server is not at Q_L , type- H' customers arrive with intensity λ_H , otherwise there are no type- H' customers arriving. In addition, type- H' customers have service requirement B_H , and type- L' customers have service requirement $B_{L'}$ with LST $\widetilde{B}_{L'}(s)$ defined in (9). Hence the traffic load of Q_L equals $\rho_{L'} = \frac{\rho_L}{1-\rho_H}$ and the traffic load of Q_H equals $\rho_{H'} = \rho_1 - \rho_{L'} = \frac{(1-\rho_1)\rho_H}{1-\rho_H}$.

Obviously, the non-priority model has the same joint queue lengths at visit beginning and visit ending of Q_2 as the original model. Besides, $\widetilde{V}_{b_i}(\cdot, \cdot, \cdot)$ and $\widetilde{V}_{c_i}(\cdot, \cdot, \cdot)$, the PGF of the joint queue lengths at visit beginning and visit ending of Q_i ($i = H, L$) can be expressed by:

$$\widetilde{V}_{b_H}(z_H, z_L, z_2) = \widetilde{V}_{b_1}(z_H, z_L, z_2) = \widetilde{V}_{b_1}\left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1}, z_2\right), \quad (13)$$

$$\widetilde{V}_{b_L}(z_H, z_L, z_2) = \widetilde{V}_{c_H}(z_H, z_L, z_2) = \widetilde{V}_{b_1}(h_H(z_L, z_2), z_L, z_2), \quad (14)$$

where $h_H(z_L, z_2) = \widetilde{\theta}_H(\lambda_L(1 - z_L) + \lambda_2(1 - z_2))$.

Let L_i denote the number of type- i customers at an arbitrary time ($i = H, L, 2$) in the original model. For writing convenience, we take the notations \mathbf{z} short for

(z_H, z_L, z_2) and $\lambda(\mathbf{z}) = \lambda_H(1 - z_H) + \lambda_L(1 - z_L) + \lambda_2(1 - z_2)$. By conditioning on the location P of the server ($P \in \{V_i : i = H, L, 2\} \cup \{S_{ij} : i, j = 1, 2\}$) in the equivalent model, we circumvent the PGFs of the joint number of customers at an arbitrary time, denoted by $\tilde{L}(\mathbf{z})$. Let $L_i^{(V_j)}$ and $L_i^{(S_{jk})}$ denote the number of type- i customers at an arbitrary time during V_j ($j = H, L, 2$) and S_{jk} ($j, k = 1, 2$) respectively. Then

$$\tilde{L}(\mathbf{z}) = \sum_{i=H,L} \rho_i \tilde{L}^{(V_i)}(\mathbf{z}) + \rho_2 \tilde{L}^{(V_2)}(\mathbf{z}) + \frac{1-\rho}{\sigma} \sum_{i=1}^2 \sum_{j=1}^2 r_{ij} \pi_j \mathbb{E} S_{ij} \tilde{L}^{(S_{ij})}(\mathbf{z}). \quad (15)$$

Let $\tilde{S}_b^{(V_i)}(z)$ denote the joint number of customers at a service beginning of Q_i and X^{past} denote the elapsed life of X . Then

$$\begin{aligned} \tilde{L}^{(V_H)}(\mathbf{z}) &= \tilde{S}_b^{(V_H)}(\mathbf{z}) \tilde{B}_H^{past}(\lambda(\mathbf{z})) \\ &= \frac{\gamma_{H'} z_H [\tilde{V}_{b_H}(\mathbf{z}) - \tilde{V}_{c_H}(\mathbf{z})]}{z_H - \tilde{B}_H(\lambda(\mathbf{z}))} \frac{1 - \tilde{B}_H(\lambda(\mathbf{z}))}{\lambda(\mathbf{z}) \mathbb{E} B_H} \\ &= \frac{\pi_1(1-\rho)}{\rho_{H'} \sigma} \frac{z_H [\tilde{V}_{b_H}(\mathbf{z}) - \tilde{V}_{b_H}(h_H(\mathbf{z}), z_L, z_2)]}{z_H - \tilde{B}_H(\lambda(\mathbf{z}))} \frac{1 - \tilde{B}_H(\lambda(\mathbf{z}))}{\lambda(\mathbf{z})}, \end{aligned} \quad (16)$$

where the last equation follows from $\gamma_{H'} = \frac{1-\rho}{1-\rho_L} \frac{\pi_1}{\lambda_H \sigma}$ since there are no arrivals of type- H customers during V_L .

Along the same argument, we get

$$\tilde{L}^{(V_2)}(\mathbf{z}) = \frac{\pi_2(1-\rho)}{\rho_2 \sigma} \frac{z_2 [\tilde{V}_{b_2}(\mathbf{z}) - \tilde{V}_{b_2}(h_2(z_H, z_L, \mathbf{z}))]}{z_2 - \tilde{B}_2(\lambda(\mathbf{z}))} \frac{1 - \tilde{B}_2(\lambda(\mathbf{z}))}{\lambda(\mathbf{z})}. \quad (17)$$

Let $\tilde{L}^{(V_{L'})}(\mathbf{z})$ denote the PGF of the joint number of customers at an arbitrary time during $B_{L'}$. Then for $\tilde{L}^{(V_L)}(\mathbf{z})$, using the same method yields

$$\begin{aligned} \tilde{L}^{(V_L)}(\mathbf{z}) &= \tilde{S}_b^{(V_L)}(\mathbf{z}) \tilde{L}^{(V_{L'})}(\mathbf{z}) \\ &= \frac{\pi_1(1-\rho)}{\lambda_L \sigma} \frac{z_L [\tilde{V}_{b_L}(\mathbf{z}) - \tilde{V}_{b_L}(1, h_{L'}(z_2), z_2)]}{z_L - \tilde{B}_{L'}(\lambda(\mathbf{z}))} \tilde{L}^{(V_{L'})}(\mathbf{z}), \end{aligned} \quad (18)$$

where $h_{L'}(z_2) = \tilde{\theta}_{L'}(\lambda_2(1 - z_2))$ and $\theta_{L'}$ is a busy time of an $M/G/1$ queue with arriving rate λ_L and service requirement $B_{L'}$.

However, $\tilde{L}^{(V_{L'})}(\mathbf{z}) \neq \tilde{B}_{L'}^{past}(\lambda(\mathbf{z}))$, since the tagged time during a service time $B_{L'}$ depends on the number of arriving type- H descendants of the type- L customer. To solve $\tilde{L}^{(V_{L'})}(\mathbf{z})$, we introduce a cyclic polling system with three queues: $Q_{H^*}, Q_{L^*}, Q_{2^*}$ (in cyclic order). Customers arrive at Q_{i^*} according to a Poisson Process with intensity λ_i . We refer the customers at Q_{i^*} as type- i^* customers. Type- H^* customer has service requirement B_H and other customers has

infinitely small service requirement, which means that once their service begins, it can be completed immediately. It is assumed that there are no switch-over times while Q_{H^*} has a setup time B_L . It is easy to see that $\tilde{L}^{(V_L)}(\mathbf{z})$ is just the PGF of the joint queue length at an arbitrary time of this system. Hence, we have

$$\begin{aligned} \tilde{L}^{(V_L)}(\mathbf{z}) = & \rho_H \frac{1 - \rho_H}{\rho_H} \frac{z_H \left[\tilde{B}_L(\lambda(\mathbf{z})) - \tilde{B}_L(\lambda(h_H(\mathbf{z}), z_L, z_2)) \right]}{z_H - \tilde{B}_H(\lambda(\mathbf{z}))} \frac{1 - \tilde{B}_H(\lambda(\mathbf{z}))}{\lambda(\mathbf{z}) \mathbb{E} B_H} \\ & + (1 - \rho_H) \frac{1 - \tilde{B}_L(\lambda(\mathbf{z}))}{\lambda(\mathbf{z}) \mathbb{E} B_L} \end{aligned} \quad (19)$$

Substituting (19) into (18) yields the expression of $\tilde{L}^{(V_L)}(\mathbf{z})$. Besides, $L_i^{(S_{jk})}$ can be expressed as

$$\tilde{L}^{(S_{1j})}(\mathbf{z}) = \tilde{V}_{b_1}(\tilde{K}_{12}(z_2), z_2) \frac{1 - M_{1j}(\mathbf{z})}{\lambda(\mathbf{z}) \mathbb{E} S_{1j}}, \quad (20)$$

$$\tilde{L}^{(S_{2j})}(\mathbf{z}) = \tilde{V}_{b_2} \left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1}, \tilde{K}_{21} \left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1} \right) \right) \frac{1 - M_{2j}(\mathbf{z})}{\lambda(\mathbf{z}) \mathbb{E} S_{2j}}. \quad (21)$$

Then the PGF of the joint number of customers at an arbitrary time can be computed by combining (16)-(21) with (15).

5. Heavy traffic asymptotics

In this section, we increase $\Lambda = \lambda_1 + \lambda_2 + \lambda_3$ to make the traffic load $\rho \rightarrow 1$ while keeping the ratios $\lambda_H : \lambda_L : \lambda_2$ fixed. When $\rho \rightarrow 1$, the queue lengths tend to infinity. Hence, we are devoted to studying the distributions of the scaled delay $\mathcal{W}_i = (1 - \rho)W_i, i = H, L, 2$. For the Markovian priority polling system, the DSA can be the first choice. Van der Mei [15, 16] introduced this technique to give all moments and the distributions of the delay in the heavy traffic. Dorsman also applied this technique to a non-priority Markovian polling system in [8]. Throughout the remainder paper, we introduce a notation \hat{x} such that $\hat{x} = \frac{x}{\rho}$, for example, $\hat{\rho}_H = \frac{\rho_H}{\rho}$.

For preparations, we introduce two Lemmas first. The first one is the Theorem of Method of Moments introduced in [16], which is applied to prove a gamma distributed limiting random variable.

Lemma 1. *Let Γ be a gamma-distributed random variable with shape parameter $\alpha + 1$ and scale parameter μ . Let $\{Y_n\}$ be a sequence of random variables with finite moments, satisfying*

$$\lim_{n \rightarrow \infty} \mathbb{E} Y_n^k = \mathbb{E} \Gamma^k, \quad k = 1, 2, \dots$$

Then $Y_n \xrightarrow{d} \Gamma$.

Lemma 2 is devoted to the expressions of the k th derivative $[f(g)]^{(k)}(x)$ of a general composite function $f(g(x))$, which is Theorem 1 presented in [17].

Lemma 2. *For $k = 1, 2, \dots$, the k th derivative of composite function $f(g(x))$ is given by*

$$[f(g)]^{(k)}(x) = \sum_{\mathbf{m}^{(k)} \in S_k} c_k(\mathbf{m}^{(k)}) f^{(l_k)}(g(x)) \prod_{i=1}^k (g^{(i)}(x))^{m_i}, l_k = \sum_{j=1}^k m_j,$$

where

$$S_k := \left\{ \mathbf{m}^{(k)} = (m_1, \dots, m_k) : m_j \text{ are non-negative integers with } \sum_{j=1}^k j m_j = k \right\},$$

and $c_k(\mathbf{m}^{(k)})$ can be calculated in the following recursive way: $c_1(1) = 1$ and

$$\begin{aligned} c_k(\mathbf{m}^{(k)}) &= c_{k-1}(m_1 - 1, m_2, \dots, m_{k-1}) I_{\{m_1 > 0\}} \\ &\quad + \sum_{j=1}^{k-1} (m_j + 1) c_{k-1}(m_1, \dots, m_{j-1}, m_j + 1, m_{j+1} - 1, m_{j+2}, \dots, m_{k-1}) \\ &\quad \times I_{\{m_{j+1} > 0\}}, \quad k = 2, 3, \dots \end{aligned}$$

Applying Lemma 1 and Lemma 2 to the DSA expressions (5) of G_1 , we obtain the following Proposition, whose proof can be found in [8].

Proposition 1. *As $\rho \rightarrow 1$, we have*

$$\lim_{\rho \rightarrow 1} \mathbb{E}[(1 - \rho)^k G_1^k] = \frac{\hat{\lambda}_1^k \prod_{j=0}^{k-1} (\alpha + j)}{\nu_1^k}, \quad k = 1, 2, \dots,$$

where

$$\begin{aligned} \alpha &= \frac{2\hat{\rho}_1\hat{\rho}_2\mathbb{E}S^{tot}}{\hat{\lambda}_H\mathbb{E}B_H^2 + \hat{\lambda}_L\mathbb{E}B_L^2 + \hat{\lambda}_2\mathbb{E}B_2^2}, \\ \nu_1 &= \frac{2\hat{\rho}_1}{\hat{\lambda}_H\mathbb{E}B_H^2 + \hat{\lambda}_L\mathbb{E}B_L^2 + \hat{\lambda}_2\mathbb{E}B_2^2}, \\ \mathbb{E}S^{tot} &= \frac{p_1}{1 - p_1} \mathbb{E}S_{11} + \mathbb{E}S_{12} + \mathbb{E}S_{21} + \frac{p_2}{1 - p_2} \mathbb{E}S_{22} = \frac{2 - p_1 - p_2}{(1 - p_1)(1 - p_2)} \sigma. \end{aligned}$$

Further, the random variable $(1 - \rho)G_1$ converges in distribution to a Gamma-distributed random variable with shape parameter $\alpha + 1$ and scale parameter $\frac{\nu_1}{\lambda_1}$ respectively.

Based on this Proposition, we can further get the following equations:

Proposition 2. *As $\rho \rightarrow 1$, we have the limiting equations:*

$$\lim_{\rho \rightarrow 1} \tilde{G}_1(1 - \frac{(1 - \rho)s}{\lambda_1}) = \left(\frac{\nu_1}{\mu_1 + s} \right)^\alpha, \quad (22)$$

$$\lim_{\rho \rightarrow 1} \tilde{G}_1\left(\frac{\lambda_H}{\lambda_1} \tilde{\theta}_H((1 - \rho)s) + \frac{\lambda_L}{\lambda_1} (1 - \frac{(1 - \rho)s}{\lambda_L})\right) = \left(\frac{(1 - \hat{\rho}_H)\nu_1}{(1 - \hat{\rho}_H)\nu_1 + s} \right)^\alpha. \quad (23)$$

Proof. The proof of the first equation can be referred to [8]. For brevity, we only give the proof of the second equation.

Note that there exists a single random variable Y that satisfies

$$\tilde{Y}(s) = \mathbb{E}e^{-sY} = \tilde{G}_1\left(\frac{\lambda_H}{\lambda_1}\tilde{\theta}_H((1-\rho)s) + \frac{\lambda_L}{\lambda_1}\left(1 - \frac{(1-\rho)s}{\lambda_L}\right)\right) = f(g(s)),$$

where $f(z) = \tilde{G}_1(z)$ and $g(s) = \frac{\lambda_H}{\lambda_1}\tilde{\theta}_H(s) + \frac{\lambda_L}{\lambda_1}\left(1 - \frac{s}{\lambda_L}\right)$. It is easy to obtain

$$\begin{aligned} f^{(k)}(1) &= \mathbb{E}G_1^k, \quad k = 1, 2, \dots, \\ g^{(1)}(0) &= -\frac{\lambda_H}{\lambda_1}(1-\rho)\mathbb{E}\theta_H - \frac{1-\rho}{\lambda_1} = -\frac{1-\rho}{\lambda_1(1-\rho_H)}, \\ g^{(k)}(0) &= (-1)^k \frac{\lambda_H}{\lambda_1}(1-\rho)^k \mathbb{E}\theta_H^k, \quad k = 2, 3, \dots \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}Y^k &= (-1)^k [f(g)]^{(k)}(0) \\ &= \sum_{\mathbf{m}^{(k)} \in S_k} c_k(\mathbf{m}^{(k)}) \mathbb{E}G_1^{l_k} \left(\frac{1-\rho}{\lambda_1(1-\rho_H)} \right)^{m_1} \prod_{i=2}^k \left(\frac{\lambda_H}{\lambda_1}(1-\rho)^i \mathbb{E}\theta_H^i \right)^{m_i} \\ &= \sum_{\mathbf{m}^{(k)} \in S_k} (1-\rho)^k c_k(\mathbf{m}^{(k)}) \mathbb{E}G_1^{l_k} \left(\frac{1}{\lambda_1(1-\rho_H)} \right)^{m_1} \prod_{i=2}^k \left(\frac{\lambda_H}{\lambda_1} \mathbb{E}\theta_H^i \right)^{m_i} \\ &\quad \prod_{i=2}^k \left(\frac{\lambda_H}{\lambda_1} \mathbb{E}\theta_H^i \right)^{m_i}. \end{aligned}$$

Since $l_k = k$ holds iff $\mathbf{m}^{(k)} = (k, 0, \dots, 0)$. Hence,

$$\lim_{\rho \rightarrow 1} \mathbb{E}Y^k = \lim_{\rho \rightarrow 1} [(1-\rho)^k \mathbb{E}G_1^k] \left(\frac{1}{\lambda_1(1-\rho_H)} \right)^k = \frac{\prod_{j=0}^{k-1} (\alpha + j)}{[\nu_1(1-\hat{\rho}_H)]^k},$$

where the second equation follows from Proposition 1. By Lemma 1, equation (23) holds, which concludes the proof. \square

Now we present the main theorem to give the limiting distributions of the scaled delay.

Theorem 3. *For $0 \leq p_1, p_2 < 1$, the LSTs of the limiting scaled waiting time distributions are given by*

$$\lim_{\rho \rightarrow 1} \widetilde{\mathcal{W}}_H(s) = (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1) \mathbb{E}S^{tot}} \left[1 - \left(\frac{\nu_1}{\nu_1 + s} \right)^\alpha \right] + \hat{\rho}_{L'}, \quad (24)$$

$$\lim_{\rho \rightarrow 1} \widetilde{\mathcal{W}}_L(s) = \frac{1 - \hat{\rho}_H}{s(1 - \hat{\rho}_1) \mathbb{E}S^{tot}} \left[1 - \left(\frac{(1 - \hat{\rho}_H)\nu_1}{(1 - \hat{\rho}_H)\nu_1 + s} \right)^\alpha \right], \quad (25)$$

$$\lim_{\rho \rightarrow 1} \widetilde{\mathcal{W}}_2(s) = \frac{1}{s(1 - \hat{\rho}_2) \mathbb{E}S^{tot}} \left[1 - \left(\frac{\nu_2}{\nu_2 + s} \right)^\alpha \right], \quad (26)$$

where α , ν_1 and $\mathbb{E}S^{tot}$ are defined in Proposition 1 and ν_2 is defined in the same way as ν_1 by $\nu_2 = \frac{2\hat{\rho}_2}{\hat{\lambda}_H \mathbb{E}B_H^2 + \hat{\lambda}_L \mathbb{E}B_L^2 + \hat{\lambda}_2 \mathbb{E}B_2^2}$. Equivalently,

$$\begin{aligned}\lim_{\rho \rightarrow 1} Pr(\mathcal{W}_H \leq t) &= (1 - \hat{\rho}_{L'}) Pr(UI_H \leq t) + \hat{\rho}_{L'}, \\ \lim_{\rho \rightarrow 1} Pr(\mathcal{W}_L \leq t) &= Pr(UI_L \leq t), \\ \lim_{\rho \rightarrow 1} Pr(\mathcal{W}_2 \leq t) &= Pr(UI_2 \leq t),\end{aligned}$$

where U is a uniformly $[0, 1]$ distributed random variable and I_i is a Gamma distributed random variable with shape parameter $\alpha + 1$ and scaled parameter ω_i ($i = H, L, 2$) with

$$\omega_H = \nu_1, \quad \omega_L = (1 - \hat{\rho}_H)\nu_1, \quad \omega_2 = \nu_2.$$

Besides, U and I_i are mutually independent ($i = H, L, 2$).

Proof. Rewriting equation (10) by (3) leads to an alternative expression

$$\begin{aligned}\widetilde{W}_H(s) &= \frac{\pi_1(1 - \rho) \left[1 - r_{11} \widetilde{M}_{11}(1 - \frac{s}{\lambda_1}, 1) - r_{21} \widetilde{G}_1(1 - \frac{s}{\lambda_1}) \right]}{\sigma[s - \lambda_H(1 - \widetilde{B}_H(s))]} \\ &\quad + \frac{\rho_L \left[1 - \widetilde{B}_L(s) \right]}{[s - \lambda_H(1 - \widetilde{B}_H(s))] \mathbb{E}B_L},\end{aligned}\tag{27}$$

Taking the limit of (27) leads to

$$\begin{aligned}\lim_{\rho \rightarrow 1} \widetilde{W}_H(s) &= \lim_{\rho \rightarrow 1} \widetilde{W}_H((1 - \rho)s) \\ &= \lim_{\rho \rightarrow 1} \frac{\pi_1(1 - \rho)}{\sigma[(1 - \rho)s - \lambda_H(1 - \widetilde{B}_H((1 - \rho)s))]} \\ &\quad \times \lim_{\rho \rightarrow 1} \left[1 - r_{11} \widetilde{M}_{11}(1 - \frac{(1 - \rho)s}{\lambda_1}, 1) - r_{21} \widetilde{G}_1(1 - \frac{(1 - \rho)s}{\lambda_1}) \right] \\ &\quad + \lim_{\rho \rightarrow 1} \frac{(1 - \rho)\rho_L}{[(1 - \rho)s - \lambda_H(1 - \widetilde{B}_H((1 - \rho)s))] \mathbb{E}B_L} \lim_{\rho \rightarrow 1} \frac{1 - \widetilde{B}_L((1 - \rho)s)}{\mathbb{E}B_L(1 - \rho)} \\ &= \frac{\pi_1}{\sigma s(1 - \hat{\rho}_H)} \left[1 - r_{11} - r_{21} \left(\frac{\nu_1}{\nu_1 + s} \right)^\alpha \right] + \frac{\rho_L}{1 - \rho_H},\end{aligned}$$

where the last equation follows from (22). Since $\frac{\pi_1}{\sigma} = \frac{1}{r_{21} \mathbb{E}S^{tot}}$, the above equation is equivalent to (24).

Rewriting equation (11) by (3) yields

$$\begin{aligned}\widetilde{W}_L(s) &= \frac{\pi_1(1 - \rho)}{\sigma[s - \lambda_L(1 - \widetilde{B}'_L(s))]} \left[1 - r_{11} \widetilde{M}_{11} \left(\frac{\lambda_H}{\lambda_1} \widetilde{\theta}_H(s) + \frac{\lambda_L}{\lambda_1} (1 - \frac{s}{\lambda_L}), 1 \right) \right. \\ &\quad \left. - r_{21} \widetilde{G}_1 \left(\frac{\lambda_H}{\lambda_1} \widetilde{\theta}_H(s) + \frac{\lambda_L}{\lambda_1} (1 - \frac{s}{\lambda_L}) \right) \right].\end{aligned}\tag{28}$$

Using Proposition 2 to (28), equations (25) can be proved in the similar way as well as (26). The proofs are omitted here. \square

6. Asymptotics with large switch-over times

In this section, we consider the limiting distributions of the scaled delay when the average total switch-over time $\sigma \rightarrow \infty$, which is equivalent to $\mathbb{E}S^{tot} \rightarrow \infty$. For convenience, we take $r = \mathbb{E}S^{tot}$ for short.

6.1. Asymptotics for deterministic switch-over times

In this subsection, we consider the case of the deterministic switch-over times. In this case, (6) becomes

$$\tilde{H}_1(z) = e^{-r_1 \sum_{c=0}^{\infty} [\lambda_1(1-\tilde{A}_{1,c-1}(z)) + \lambda_2(1-\tilde{A}_{2,c}(z))] - r_2 \sum_{c=0}^{\infty} [\lambda_1(1-\tilde{A}_{1,c-1}(z)) + \lambda_2(1-\tilde{A}_{2,c-1}(z))]}, \quad (29)$$

where $r_1 = \mathbb{E}R_1 = \mathbb{E}S_{12} + \frac{p_2}{1-p_2}\mathbb{E}S_{22}$ and $r_2 = \mathbb{E}R_2 = \frac{p_1}{1-p_1}\mathbb{E}S_{11} + \mathbb{E}S_{21}$. Let $\alpha_{i,c}^{(k)}$ denote the k th moment of $\tilde{A}_{i,c}(z)$ ($k = 1, 2, \dots$). By differentiating $\tilde{A}_{i,c}(z)$, we can compute $\alpha_{i,c}^{(1)}$ in the recursive way:

$$\alpha_{1,c}^{(1)} = \mathbb{E}\theta_1 \lambda_2 \alpha_{2,c}^{(1)}, \quad \alpha_{2,c}^{(1)} = \mathbb{E}\theta_2 \lambda_1 \alpha_{1,c-1}^{(1)}. \quad (30)$$

Rewriting equation (30) to a matrix form yields

$$\begin{pmatrix} \alpha_{1,c}^{(1)} \\ \alpha_{2,c}^{(1)} \end{pmatrix} = \begin{pmatrix} \lambda_1 \lambda_2 \mathbb{E}\theta_1 \mathbb{E}\theta_2 & 0 \\ \lambda_1 \mathbb{E}\theta_2 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1,c-1}^{(1)} \\ \alpha_{2,c-1}^{(1)} \end{pmatrix} = \begin{pmatrix} \lambda_1 \lambda_2 \mathbb{E}\theta_1 \mathbb{E}\theta_2 & 0 \\ \lambda_1 \mathbb{E}\theta_2 & 0 \end{pmatrix}^{c+1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \quad (31)$$

By differentiating equation (29), after substituting (31), we obtain

$$\begin{aligned} \mathbb{E}H_1 &= r_1 \sum_{c=0}^{\infty} \left(\lambda_1 \alpha_{1,c-1}^{(1)} + \lambda_2 \alpha_{2,c}^{(1)} \right) + r_2 \sum_{c=0}^{\infty} \left(\lambda_1 \alpha_{1,c-1}^{(1)} + \lambda_2 \alpha_{2,c-1}^{(1)} \right) \\ &= (r_1 + r_2) \sum_{c=-1}^{\infty} \left(\lambda_1 \alpha_{1,c}^{(1)} + \lambda_2 \alpha_{2,c}^{(1)} \right) \\ &= r \sum_{c=-1}^{\infty} (\lambda_1, \lambda_2) \begin{pmatrix} \alpha_{1,c}^{(1)} \\ \alpha_{2,c}^{(1)} \end{pmatrix} \\ &= r \left[\lambda_1 + \sum_{c=0}^{\infty} (\lambda_1, \lambda_2) \begin{pmatrix} \lambda_1 \lambda_2 \mathbb{E}\theta_1 \mathbb{E}\theta_2 & 0 \\ \lambda_1 \mathbb{E}\theta_2 & 0 \end{pmatrix}^{c+1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right] \\ &= r \left[\lambda_1 + \sum_{c=0}^{\infty} \lambda_1 \lambda_2 \mathbb{E}\theta_2 (1 + \lambda_1 \mathbb{E}\theta_1) (\lambda_1 \lambda_2 \mathbb{E}\theta_1 \mathbb{E}\theta_2)^c \right] \\ &= \frac{\lambda_1 (1 - \rho_1) r}{1 - \rho}. \end{aligned} \quad (32)$$

Introduce $\tilde{G}_{L'}(z)$ such that

$$\tilde{G}_{L'}(z) = \tilde{G}_1 \left(\frac{\lambda_H}{\lambda_1} \tilde{\theta}_H(\lambda_L(1-z)) + \frac{\lambda_L}{\lambda_1} z \right), \quad (33)$$

and a random variable K that satisfies

$$\begin{aligned} \tilde{K}(z) = & e^{-r_1 \sum_{c=0}^{\infty} [\lambda_H(1 - \tilde{B}_{H,c-1}(z)) + \lambda_L(1 - \tilde{B}_{L,c-1}(z)) + \lambda_2(1 - \tilde{B}_{2,c}(z))]} \\ & \times e^{-r_2 \sum_{c=0}^{\infty} [\lambda_H(1 - \tilde{A}_{H,c-1}(z)) + \lambda_L(1 - \tilde{B}_{L,c-1}(z)) + \lambda_2(1 - \tilde{B}_{2,c-1}(z))]}, \end{aligned} \quad (34)$$

where

$$\begin{aligned} \tilde{B}_{H,c}(z) &= \tilde{\theta}_H(\lambda_L(1 - \tilde{B}_{L,c}(z)) + \lambda_2(1 - \tilde{B}_{2,c}(z))), & \tilde{B}_{H,-1}(z) &= \tilde{\theta}_H(\lambda_L(1 - z)), \\ \tilde{B}_{L,c}(z) &= \tilde{\theta}'_L(\lambda_2(1 - \tilde{B}_{2,c}(z))), & \tilde{B}_{L,-1}(z) &= z \\ \tilde{B}_{2,c}(z) &= \tilde{\theta}_2(\lambda_H(1 - \tilde{B}_{H,c-1}(z)) + \lambda_L(1 - \tilde{B}_{L,c-1}(z))), & \tilde{B}_{2,-1}(z) &= 1. \end{aligned}$$

With some efforts, we can prove

$$\tilde{G}_{L'}(z) = \tilde{K}(z) \frac{1 - r_{11} \tilde{M}_{11}(\tilde{\theta}_H(\lambda_L(1 - z)), z, 1)}{r_{21}}. \quad (35)$$

Combining equation (5), (33) with (35) leads to

$$\tilde{K}(z) = \tilde{H}_1\left(\frac{\lambda_H}{\lambda_1} \tilde{\theta}_H(\lambda_L(1 - z)) + \frac{\lambda_L}{\lambda_1} z\right). \quad (36)$$

Differentiating (36), we obtain the first moment of K :

$$\mathbb{E}K = \mathbb{E}H_1\left(\frac{\lambda_H}{\lambda_1} \mathbb{E}\theta_H \lambda_L + \frac{\lambda_L}{\lambda_1}\right) = \frac{\lambda_L(1 - \rho_{L'})r}{1 - \rho}.$$

Applying the Strong Law of Large Numbers for Renewal Reward Processes to the DSA expressions (29) and (34) yields $\frac{H_1}{r} \rightarrow \frac{\mathbb{E}H_1}{r}$ and $\frac{K}{r} \rightarrow \frac{\mathbb{E}K}{r}$ when r tends to infinity. Here we just present the conclusions. Interested readers can refer to Theorem 1 in [15] for more details.

Proposition 3. *When $r \rightarrow \infty$, we have*

$$\frac{H_1}{r} \xrightarrow{a.s.} \frac{\lambda_1(1 - \rho_1)}{1 - \rho}, \quad \frac{K}{r} \xrightarrow{a.s.} \frac{\lambda_L(1 - \rho_{L'})}{1 - \rho}.$$

Theorem 4. *For $0 \leq p_1, p_2 < 1$, when the switch-over times tend to infinity, we have*

$$\lim_{r \rightarrow \infty} \mathbb{E}[e^{-s \frac{W_H}{r}}] = (1 - \rho_{L'}) \frac{(1 - \rho)}{s(1 - \rho_1)} \left[1 - e^{-\frac{1 - \rho_1}{1 - \rho} s}\right] + \rho_{L'}, \quad (37)$$

$$\lim_{r \rightarrow \infty} \mathbb{E}[e^{-s \frac{W_L}{r}}] = \frac{(1 - \rho)}{s(1 - \rho_{L'})} \left[1 - e^{-\frac{1 - \rho_{L'}}{1 - \rho} s}\right], \quad (38)$$

$$\lim_{r \rightarrow \infty} \mathbb{E}[e^{-s \frac{W_2}{r}}] = \frac{(1 - \rho)}{s(1 - \rho_2)} \left[1 - e^{-\frac{1 - \rho_2}{1 - \rho} s}\right]. \quad (39)$$

Equivalently,

$$\lim_{r \rightarrow \infty} Pr\left(\frac{W_H}{r} \leq t\right) = (1 - \rho_{L'}) Pr(U_H \leq t) + \rho_{L'},$$

$$\lim_{r \rightarrow \infty} \Pr\left(\frac{W_L}{r} \leq t\right) = \Pr(U_L \leq t),$$

$$\lim_{r \rightarrow \infty} \Pr\left(\frac{W_2}{r} \leq t\right) = \Pr(U_2 \leq t),$$

where U_i ($i = H, L, 2$) is a uniformly $[0, u_i]$ distributed random variable with

$$u_H = \frac{1 - \rho_1}{1 - \rho}, \quad u_L = \frac{1 - \rho_{L'}}{1 - \rho}, \quad u_H = \frac{1 - \rho_2}{1 - \rho}.$$

Proof. From Propostion 3, we have

$$\lim_{r \rightarrow \infty} \tilde{H}_1\left(1 - \frac{s}{\lambda_1 r}\right) = \lim_{r \rightarrow \infty} \mathbb{E}\left(1 - \frac{s}{\lambda_1 r}\right)^{-\frac{\lambda_1 r}{s} \frac{H_1}{r}}\left(-\frac{s}{\lambda_1}\right) = \mathbb{E}e^{-\frac{H_1}{r} \frac{s}{\lambda_1}} = e^{-\frac{1-\rho_1}{1-\rho} s}, \quad (40)$$

The same way gives

$$\lim_{r \rightarrow \infty} \tilde{K}\left(1 - \frac{s}{\lambda_L r}\right) = e^{-\frac{(1-\rho_{L'})}{1-\rho} s}. \quad (41)$$

By the expression (27) of $\tilde{W}_H(s)$, we have

$$\begin{aligned} \lim_{r \rightarrow \infty} \mathbb{E}[e^{-s \frac{W_H}{r}}] &= \lim_{r \rightarrow \infty} \frac{\pi_1(1 - \rho) \left[1 - r_{11} \tilde{M}_{11}\left(1 - \frac{s}{\lambda_1 r}, 1\right) - r_{21} \tilde{G}_1\left(1 - \frac{s}{\lambda_1 r}\right)\right]}{\sigma\left[\frac{s}{r} - \lambda_H\left(1 - \tilde{B}_H\left(\frac{s}{r}\right)\right)\right]} \\ &\quad + \lim_{r \rightarrow \infty} \frac{\rho_L(1 - \tilde{B}_L(\frac{s}{r}))}{\left[\frac{s}{r} - \lambda_H\left(1 - \tilde{B}_H(\frac{s}{r})\right)\right] \mathbb{E}B_L}, \\ &= \lim_{r \rightarrow \infty} \frac{(1 - \rho) \left[1 - r_{11} \tilde{M}_{11}\left(1 - \frac{s}{\lambda_1 r}, 1\right) - r_{21} \tilde{G}_1\left(1 - \frac{s}{\lambda_1 r}\right)\right]}{r_{21} s \left[1 - \lambda_H \frac{1 - \tilde{B}_H(\frac{s}{r})}{\frac{s}{r}}\right]} \\ &\quad + \lim_{r \rightarrow \infty} \frac{\rho_L(1 - \tilde{B}_L(\frac{s}{r}))}{\left[\frac{s}{r} - \lambda_H\left(1 - \tilde{B}_H(\frac{s}{r})\right)\right] \mathbb{E}B_L}, \\ &= \lim_{r \rightarrow \infty} \frac{(1 - \rho) \left[1 - r_{11} - r_{21} \tilde{H}_1\left(1 - \frac{s}{\lambda_1 r}\right)\right]}{r_{21} s (1 - \rho_H)} + \frac{\rho_L}{1 - \rho_H} \\ &= \frac{(1 - \rho)}{s(1 - \rho_H)} \left[1 - e^{-\frac{1-\rho_1}{1-\rho} s}\right] + \frac{\rho_L}{1 - \rho_H}. \end{aligned}$$

where the second equation follows from $\frac{\pi_1}{\sigma} = \frac{1}{r_{21} r}$, the third equation follows from equation (5) and the last equation results from equation (40), which is equivalent to (37).

By combining (27), (28), (40) and (41), (38) and (39) can be proved along the same methodology. I won't go into much details here. \square

6.2. Asymptotics for nondeterministic switch-over times under heavy traffic

The aim of the present subsection is to study the asymptotic delay for general switch-over times under the heavy traffic when the switch-over times tend to infinity, which actually is the heavy-traffic behaviors of (24)-(26).

Theorem 5.

$$\lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} \mathbb{E}[e^{-s(1-\rho)\frac{W_H}{r}}] = (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1)} [1 - e^{-(1-\hat{\rho}_1)s}] + \hat{\rho}_{L'}, \quad (42)$$

$$\lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} \mathbb{E}[e^{-s(1-\rho)\frac{W_L}{r}}] = \frac{1 - \hat{\rho}_H}{s(1 - \hat{\rho}_{L'})} \left[1 - e^{-\frac{(1-\hat{\rho}_{L'})}{1-\hat{\rho}_H}s} \right], \quad (43)$$

$$\lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} \mathbb{E}[e^{-s(1-\rho)\frac{W_2}{r}}] = \frac{1}{s(1 - \hat{\rho}_2)} [1 - e^{-(1-\hat{\rho}_2)s}]. \quad (44)$$

Equivalently,

$$\begin{aligned} \lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} Pr\left(\frac{(1-\rho)W_H}{r} \leq t\right) &= (1 - \hat{\rho}_{L'})Pr(U_{H1} \leq t) + \hat{\rho}_{L'}, \\ \lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} Pr\left(\frac{(1-\rho)W_L}{r} \leq t\right) &= Pr(U_{L1} \leq t), \\ \lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} Pr\left(\frac{(1-\rho)W_2}{r} \leq t\right) &= Pr(U_{21} \leq t), \end{aligned}$$

where U_{i1} ($i = H, L, 2$) is a uniformly $[0, u_{i1}]$ distributed random variable with

$$u_{H1} = 1 - \hat{\rho}_1, \quad u_{L1} = \frac{1 - \hat{\rho}_{L'}}{1 - \hat{\rho}_H}, \quad u_{21} = 1 - \hat{\rho}_2.$$

Proof. Since equations (42)-(44) follow the same arguments, for compactness, we only present the derivation of (42). By (24), we have

$$\begin{aligned} \lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} \mathbb{E}[e^{-s(1-\rho)\frac{W_H}{r}}] &= \lim_{r \rightarrow \infty} (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1)} \left[1 - \left(\frac{\nu_1}{\nu_1 + \frac{s}{r}} \right)^\alpha \right] + \hat{\rho}_{L'} \\ &= \lim_{r \rightarrow \infty} (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1)} \left[1 - \left(1 + \frac{s}{\nu_1 r} \right)^{-\frac{\nu_1 r}{s} \frac{\alpha_1 s}{\nu_1}} \right] + \hat{\rho}_{L'} \\ &= (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1)} \left[1 - e^{-\frac{\alpha_1 s}{\nu_1}} \right] + \hat{\rho}_{L'} \\ &= (1 - \hat{\rho}_{L'}) \frac{1}{s(1 - \hat{\rho}_1)} [1 - e^{-(1-\hat{\rho}_1)s}] + \hat{\rho}_{L'}, \end{aligned}$$

where $\alpha_1 = \frac{2\hat{\rho}_1\hat{\rho}_2}{\hat{\lambda}_H\mathbb{E}B_H^2 + \hat{\lambda}_L\mathbb{E}B_L^2 + \hat{\lambda}_2\mathbb{E}B_2^2}$ and hence, $\alpha = \alpha_1 r$ and $\frac{\alpha_1}{\nu_1} = \hat{\rho}_2 = 1 - \hat{\rho}_1$. \square

Remark 1. Compared to the non-priority model, Theorem 3-5 illustrate the following interesting phenomenons:

1. The scaled delay of a high-priority customer is a modified distribution of the scaled delay of a type-1 customer in the non-priority model.
2. The scaled delay of a low-priority customer has the same distribution as the scaled delay of a type-1 customer in the non-priority model except the scale parameter.
3. The scaled delay of a type-2 customer remains the same as that in the non-priority model.

7. Numerical examples

In this section we test the validity of the asymptotic results in Theorem 3-5, mainly by simulating the polling system with different ρ and r , and then comparing the empirical Cumulative Distribution Function (CDF) of the scaled delay and the asymptotic delay. For simplicity, we consider a polling system with exponentially distributed service time and switch-over times.

Tab. 1 Parameter values in the test of the heavy-traffic behaviors

Parameter	Considered parameter values
Traffic load	$\rho \in \{0.5, 0.8, 0.9, 0.95, 0.99\}$
Mean service time	$B = 0.85$
Probability of transitions	$p_1 = 0.4, p_2 = 0.3$
Mean switch-over times	$\sigma = 2.4$

First we consider the polling system in the heavy traffic. The parameter values of the tested polling system can be found in Tab. 1. We use Matlab to undertake simulations and each simulation runs until at least 2000 customers are served. We take down the waiting times of each customer and plot the CDF of each type of customer in Fig.1. The asymptotic distribution of the scaled delay is plotted in the same graph. Actually, the asymptotic distribution of the scaled delay is a residual life of a Gamma distribution or modified Gamma distribution, and hence, can be plotted by the Matlab function “gamcdf”. From Fig.1, it is readily showed that the approximation performs very well when ρ is close to 1.

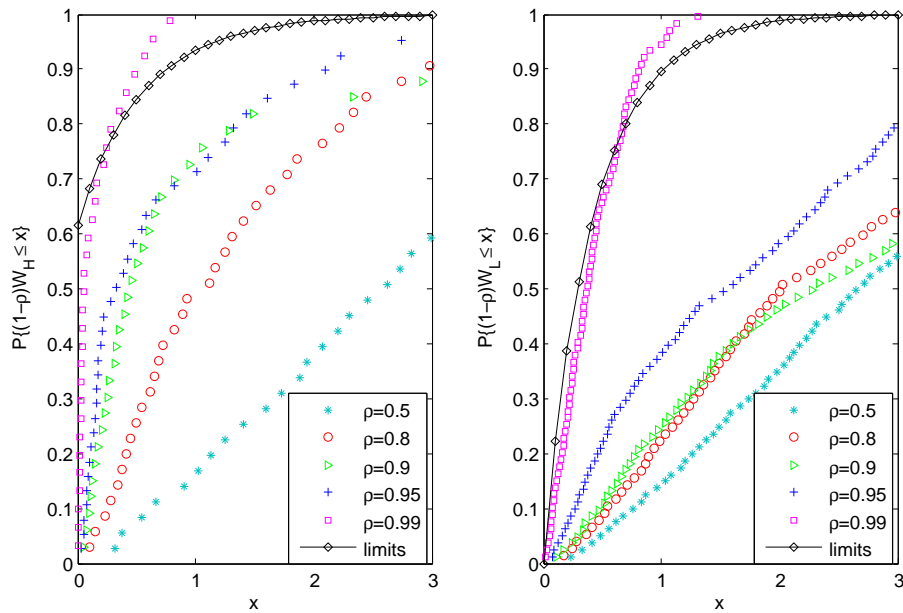


Fig. 1 The CDFs of scaled delay of type- H and type- L customers for different load

Tab. 2 Parameter values in the test under large deterministic switch-over times

Parameter	Considered parameter values
Traffic load	$\rho = 0.8$
Mean service time	$B = 0.85$
Probability of transitions	$p_1 = 0.4, p_2 = 0.3$
Mean switch-over times	$r \in \{1, 10, 50, 100, 500\}$

The scaled delay with large deterministic switch-over times converges to a uniform distribution, which is showed in Fig.2 and the parameter values of the investigated polling system is given in Tab.2. With the same polling simulation procedure, we take down the waiting times of at least 5000 customers and plot their CDFs with Matlab function “cdfplot”. In Fig. 2, the endpoints of the CDFs in the simulation converge to the limits while other points haven't converged completely. With regard to the running time and the tolerance error, the convergence trend can not be denied in Fig.2.

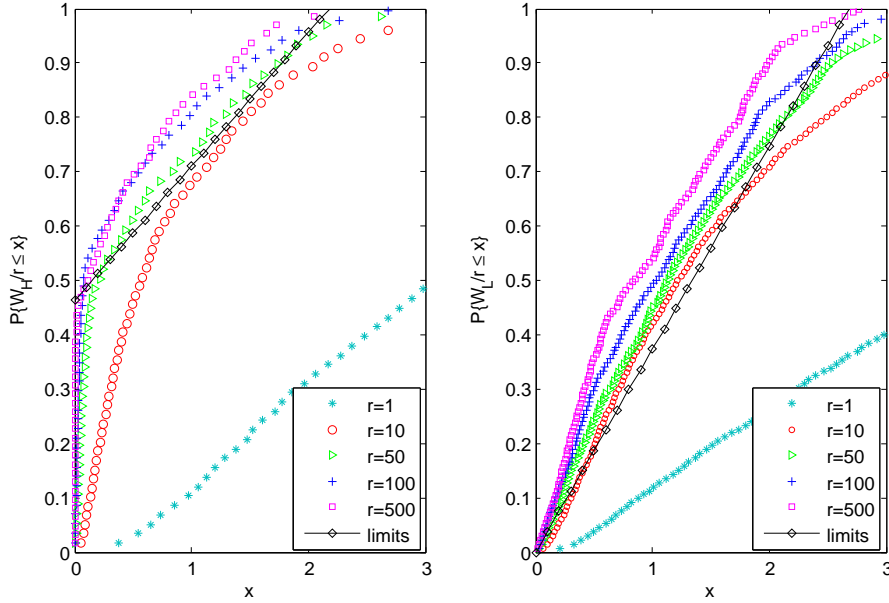


Fig. 2 The CDFs of scaled delay of type- H and type- L customers under different deterministic switch-over times

As for the case of general switch-over times, we consider the convergence process $\lim_{r \rightarrow \infty} \lim_{\rho \rightarrow 1} \frac{(1-\rho)W_i}{r}$. We undertake a simulation work with $\rho = 0.99$ and exponentially distributed switch-over times with the same means as the case of the deterministic switch-over times. 40000 customers are served in the simulation procedure. However, the convergence process performs not well. One reason is for the superposition of two limits which enlarges the error of simulation, and the other is that the utilization of two times of scale operations leads to especially

very long time required for the simulation to run. To test the validity of Theorem 5, the key is to exploit a new and effective simulation method.

8. Conclusions

In this paper, we have introduced the priority policy into a 2-queue Markovian polling system and presented some performance measures like the joint number of customers and waiting times. Moreover, with the DSA, we also investigated the limits of the scaled delays in the case of the heavy-traffic and large switch-over times. Although we have made some achievements of the Markovian priority polling system, it leaves many extension works. Here we just list some topics for further research.

1. *Mixed gated/exhaustive service* In this case, we can follow the argument in [1].
2. *Multiple priority levels* Here we can follow the argument in [2]. As for the heavy traffic behaviors and asymptotics with large switch-over times, it is an interesting subject to exploit that whether there are some beautiful conclusions using the DSA.
3. *Simulation method under large general switch-over times* Obviously, the simulation method in our paper can not work effectively for this case. No such simulation work has been published. However, this simulation can help us to exploit the asymptotics of delays with renewal arrivals and to study the limits only with large switch-over times and not in the heavy-traffic.

References

- [1] BOON, M. A. A. AND ADAN, I. J. B. F. (2009). Mixed gated/exhaustive service in a polling model with priorities. *Queueing Systems* **63**, 383–399.
- [2] BOON, M. A. A., ADAN, I. J.-B. F. AND BOXMA, O. J. (2010). A polling model with multiple priority levels. *Performance Evaluation* **67**, 468–484.
- [3] BOON, M. A. A., ADAN, I. J. B. F. AND BOXMA, O. J. (2010). A two-queue polling model with two priority levels in the first queue. *Discrete Event Dynamic Systems* **20**, 511–536.
- [4] BOON, M. A. A., VAN WIJK, A. C. C., ADAN, I. J.-B. F. AND BOXMA, O. J. (2010). A polling model with smart customers. *Queueing Systems* **66**, 239–274.

- [5] BORST, S. C. AND BOXMA, O. J. (1997). Polling models with and without switchover times. *Operations Research* **45**, 536–543.
- [6] BOXMA, O. J. AND WESTSTRATE, J. A. (1989). Waiting times in polling systems with markovian server routing. In *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*. ed. G. Stiege and J. Lie. vol. 218 of *Informatik-Fachberichte*. Springer Berlin Heidelberg pp. 89–104.
- [7] DORSMAN, J.-P. L., BORST, S. C., BOXMA, O. J. AND VLASIOU, M. (2014). Markovian polling systems with an application to wireless random-access networks. *Technical report*.
- [8] DORSMAN, J.-P. L., BOXMA, O. J. AND VAN DER MEI, R. D. (2014). On two-queue Markovian polling systems with exhaustive service. *Queueing Systems*.
- [9] FOSS, S. AND LAST, G. (1996). Stability of polling systems with exhaustive service policies and state-dependent routing. *The Annals of Applied Probability* **6**, 116–137.
- [10] FUHRMANN, S. AND COOPER, R. B. (1985). Stochastic decompositions in the $m/g/1$ queue with generalized vacations. *Operations research* **33**, 1117–1129.
- [11] KELLA, O. AND YECHIALI, U. (1988). Priorities in $M/G/1$ queue with server vacations. *Naval Research Logistics* **35**, 23–34.
- [12] LIU, Z., CHU, Y. AND WU, J. (2014). Heavy-traffic Asymptotics of Priority Polling System with Threshold Service Policy. *Submitted to Computers & Operation Research*.
- [13] LIU, Z., CHU, Y. AND WU, J. (2014). On the Three-queue Priority Polling System with Threshold Service Policy. *Submitted to Applied Mathematical Modelling*.
- [14] SRINIVASAN, M. M. (1991). Nondeterministic polling systems. *Management Science* **37**, 667–681.
- [15] VAN DER MEI, R. D. (1999). Delay in polling systems with large switch-over times. *Journal of Applied Probability* **36**, 232–243.
- [16] VAN DER MEI, R. D. (1999). Distribution of the delay in polling systems in heavy traffic. *Performance Evaluation* **38**, 133–148.
- [17] VAN DER MEI, R. D. (2000). Polling Systems with Switch-over Times under Heavy Load: Moments of the Delay. *Queueing Systems* **36**, 381–404.